Fourier - Bessel based Cepstral Coefficient Features for Text-Independent Speaker Identification

Chetana Prakash and Suryakanth V. Gangashetty

Speech and Vision Laboratory
Department of LTRC
International Institute of Technology Hyderabad - 577 032, Andra pradesh , India
email: chetana@research.iiit.ac.in, svg@iiit.ac.in

Abstract. This paper proposes the Fourier-Bessel cepstral coefficients (FBCC) as features for robust text-independent speaker identification. Fourier-Bessel (FB) expansion is used instead of Fourier transform for representing the signal in frequency domain. FB expansion can be viewed as two-dimensional Fourier transform. Change in the kernel of the transform from exponential to decaying exponentials helps in viewing the speech signal as a linear sum of decaying exponentials. For signals arising out of acoustic tubes, where the signal is subjected to many damping effects, delays in the different components of the signal is inevitable. Representing such signals using FB coefficients helps in able identification of different components present in the signal. The random non-stationary nature of speech signal is more efficiently represented by damped sinusoidal nature of basis function that is more natural for the voiced speech signal since Bessel functions have damped sinusoidal as basis function, so it is more natural choice for the representation of natural signals. Vocal tract is modeled as a set of linear acoustic tubes being cylindrical in shape can be efficiently modeled using FB expansion because Bessel functions are solutions to cylindrical wave equations. The proposed approach to speaker identification is based on FBCC features, and method employ Gaussian mixture for modeling the speaker characteristics. However, we have build the speaker models from the Fourier-Bessel features derived from the speech samples, as an alternative to Mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) for building the speaker models. An evaluation of the Gaussian mixture model is conducted on TIMIT database which consists of 630 speakers and 10 speech utterances per speaker and white noise signals of TIMIT database having various SNRs of 50, 40, 30 and 20 dB. Using the statistical model like Gaussian mixture model (GMM) and features extracted from the speech signals build a unique identity for each person who enrolled for speaker identification [1]. Estimation and Maximization algorithm is used for finding the maximum likelihood solution for a model with features, to test the later speeches against the database of all speakers who enrolled in the database. Experimental results shows that the FBCC can be used as the alternate feature for the LPCC and MFCC since it can improve the performance of the speaker identification task.

Keywords: Bessel coefficients, speaker identification, Gaussian mixer model, MFCC, FBCC

1 Introduction

Numerous measurements and signals have been proposed and investigated for use in biometric recognition systems. Among the most popular measurements are fingerprint, face, and voice. While each has pros and cons relative to accuracy and deployment, there are two main factors that have made voice a compelling biometric. In particular, speech is a natural and convenient form of input that carries the signature of the speaker. Moreover, speech is inexpensive to collect and analyze, also it is hard to mimic. Therefore, automatic speaker recognition is suitable for such applications.

The potential applications of speaker recognition systems exists when speakers are unknown and their identities are important. Access to cars, buildings, bank account and other services may be voice controlled in the future. Some existing applications use voice in conjunction with other security measures, perhaps a codeword, to provide an extra level of security. We may want to verify that the speaker we are talking to is in fact who he or she claims to be. The technology has applications to human-machine interfaces, where intelligent machines would be programmed to adapt and respond to the current user. Speech recognition systems can usefully employ speaker-recognition technology.

Speaker recognition is one area of artificial intelligence where machines performance can exceed human performance using short test utterances and a large number of speakers [2]. This is especially true for unfamiliar speakers, where the training time for humans to learn a new voice is normally very long compared to that of machines. Human performance in adverse conditions was also reviewed, where it was reported that human listeners are adept at using various cues to verify speakers in the presence of acoustic mismatch [3].

Speech is produced from a time varying vocal tract system excited by a time varying excitation source [4–6]. The resulting speech signal contains information about message, speaker, language and emotional status. For analysis and processing of speech signals, the vocal tract system is modeled as a time varying filter, and the excitation as voiced or unvoiced or plosive or combination of these types. The time varying filter characteristics capture variations in the shape of the vocal tract system in the form of resonances, anti-resonances and spectral roll-off characteristics. These filter characteristics are usually represented by spectral features for each short (10-30 ms) segment of speech, and we call these features as system features. This representation of speech has been extensively exploited for developing speaker recognition systems [7–10, 5].

Speaker recognition [11] can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. The system that we will describe is classified as text-independent speaker identification system since its task is to identify the person who speaks regardless of what is saying.

This paper approaches the speaker recognition field, an important biometric domain, providing a text independent recognition system. The most successful speech-independent recognition methods are based on vector quantization

(VQ) or Gaussian mixture model (GMM). Speaker (voice) recognition encompasses both identification and verification of speakers [12]. The basic structure of speaker identification is shown in the following Figure 1.

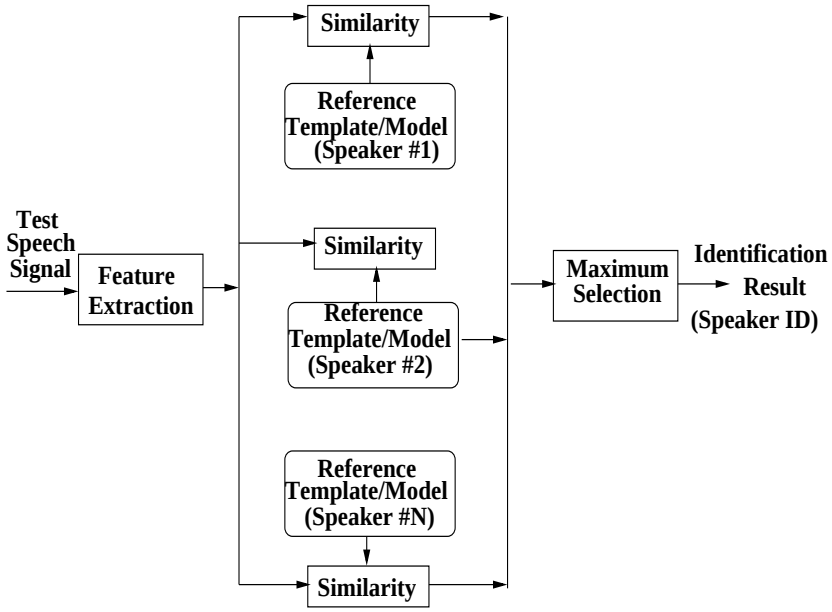

Fig. 1. *Basic flow diagram of a closed-set speaker identification system.*

In the identification, the speaker whose model best matches with the test utterance is declared as the identified speaker. The output of the system is the identity of the test speaker.

The primary task in a speaker recognition system is to extract features capable of representing the speaker information present in the speech signal. The purpose of feature extraction is to describe the acoustic properties of the speakers in the speech data. Feature extraction is the process of reducing data while retaining the classification information. The selection of the best parametric representation for acoustic data is an important task in the design of any text-independent speaker recognition system. The acoustic features should fulfill the following requirements.

- Be of low dimensionality to allow a reliable estimation of parameters of the automatic speaker recognition systems.
- Be independent of the speech and recording environment.
- High inter-speaker discriminating.
- Low intra-speaker variability.
- Robust to channel characteristics and noise.

One early problem with speaker recognition systems was to choose the right acoustic features and the right acoustic models to work with. Acoustic models were chosen to be GMM , as they are assumed to offer a good fit to the statistical nature of speech. Moreover, the acoustic models are often assumed to have diagonal covariance matrices arising the need to have speech features that are by nature uncorrelated. As with any pattern recognition task, the speech signal is first reduced to a sequence of feature vectors, Speaker recognition is often based on the premise that the speaker-specific information derived from speech utterance of an individual is characterized by a unique distribution of feature vectors. That is, the feature vectors extracted from the voice of an individual occupy a region in the feature space in a manner unique to him. For better discrimination, the distribution of the feature vectors should have high inter-speaker variability and low intra-speaker variations. Text-independent systems are based on modeling the speaker's acoustic feature space. The distribution of each speaker is determined from the feature vectors obtained from his/her speech. It is a fact that the spectrum of a signal is prone to channel characteristics and noise. Channel characteristics and noise play a prominent role in the performance of spectral feature-based systems.

The performance of ASR depends primarily upon the effectiveness of the feature vector used. Although there are no exclusively speaker distinguishing speech features, the speech spectrum has been shown to be very effective for speaker identification [7]. This is because the spectrum reflects a person's vocal tract structure, the predominant physiological factor which distinguishes one person's voice from others. LPC spectral representations, such as LPC cepstral and reflection coefficients, have been used extensively for speaker recognition; however, these model-based representations can be severely affected by noise [13]. Recent studies have found directly computed filter bank features to be more robust for noisy speech recognition.

Most state-of-art speaker recognition systems use Mel frequency cepstral coefficients (MFCC) as the acoustic features, primarily because of MFCC's superior robustness to additive noise. Bandwidth of Mel-filter and number of Mel-filters are two important parameters in the design of a Mel-filter bank. The choice of the number of filters has no specific criteria and is generally based on the type of the task as well as on the sampling frequency of the database. Stevens and Volkmann, developed the Mel-scale as a result of a study of the human auditory perception [14]. The Mel scale was used by Mermelstein and Davis to extract features from the speech signal for improved recognition performance [15].

Since all the real world services have to deal with speech coming over telephone channel, the ASR systems have to be robust to environmental variations. So Attempt made to develop an approach which can work for large amount of speech data and also which is robust to noisy conditions is proposed in this work.

The Fourier series representation employs an orthogonal set of sinusoidal functions as a basis set, while the Fourier transforms uses a complex exponential function, related to the sinusoidal through Eulers relation, as its kernel [16]. The sinusoidal functions are the periodic and ideal for representing general periodic

functions. But it may not fully match the properties of other waveforms. In particular, the random non-stationary nature of speech waveforms do not lead to the most efficient representation by sinusoidal based transformations.

In case of non-stationary signals, an aperiodic signal set may be more efficient for representation. Based on this premise, several aperiodic non-sinusoidal functions including exponentially modulated sinusoids and Bessel functions of the first kind have been used for speech analysis with varying degrees of success [17–23]. The Bessel functions has a orthogonal set of damped sinusiodal basis functions which results in an expansion termed the Fourier-Bessel (FB) series [24, 16, 25, 26]. In the present work we consider the representation of non-stationary signals using the zeroth-order Bessel functions.

The organization of the paper is as follows: The motivation for using damped sinusoidal as basis function is given in the Section 2. The use of GMMs for text-independent speaker identification is given in Section 3. The theory about the Fourier-Bessel is explained in Section 4. In Section 5 the performance evaluation of the Fourier-Bessel cepstral coefficients (FBCC) and Mel frequency cepstral coefficients (MFCC) on the TIMIT database is presented.

2 Motivation for using Damped Sinusoids as Basis Functions

In the Fourier-Bessel transformation the basis function is of damped sinusoidal nature which is more natural for the voiced speech signal. There are at least three reasons which make use of undamped sinusoids for base functions undesirable in the case of voiced-sound waveforms. The reasons are:

1. While successive pitch periods resemble each other to a considerable degree, the duration of this quasi-periodic function is limited, and thus, Fourier analysis in terms of the fundamental pitch frequency and its harmonics is not strictly applicable.
2. Because of variation of pitch and volume, successive pitch periods seldom have exactly the same waveform.
3. From the mechanism of generation of voiced sounds, it is known that a pulse-like excitation, originated by the action of the vocal cords, excites the various resonant cavities of the vocal tract and thus starts a combination of decaying oscillatory functions.

Thus, an approximation of the decaying functions by ordinary sinusoids does not appear to be too efficient[17]. Thus a compact representation of speech is possible using Bessel functions because of similarity between voiced speech and the Bessel functions. Both voiced speech and the Bessel functions exhibit quasi-periodicity and decaying amplitude with time.

3 Gaussian Mixture Speaker Model (Parametric) Approach

Parametric approaches are model-based approaches. The parameters of the model are estimated using the training feature vectors, it is assumed that the model is adequate to represent the distribution. The most widely used parametric approaches are Gaussian mixture model (GMM) and hidden Markov model (HMM) based approaches. GMM is used in speaker recognition applications as a generic probabilistic model for multivariate densities capable of representing arbitrary densities, which makes it well suited for unconstrained text-independent applications. The use of GMMs for text-independent speaker identification was first described in [27–29]. An extension of GMM-based systems to speaker verification was described and evaluated on several available speech corpora in [5, 30–34]

This section describes the form of the GMM and motivates its use as a representation of speaker identity for text-independent speaker identification. The speech analysis for extracting the Mel cepstral feature representation is done first. Next the Gaussian mixture speaker model and its parameterization are described. The use of the Gaussian mixture density for speaker identification is then motivated by two interpretations. First, the individual component Gaussian in a speaker-dependent GMM are interpreted to represent some broad acoustic classes. These acoustic classes reflect some general speaker-dependent vocal tract configuration that are useful for modeling speaker identity. Second, a Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker.

4 Fourier-Bessel Series

As similar to Fourier series, Fourier-Bessel series are another kind of infinite series expansion on a finite interval, based on Bessel function which are solutions of the differential equation

$$t^2 y^{''} + t y^{'} + (t^2 - n^2) y = 0, n > 0 \tag{1}$$

which is called Bessel's differential equation. The general solution is given by:

$$y = C_1 J_n(t) + C_2 Y_n(t) \tag{2}$$

where $J_n(t)$ is called a Bessel function of the first kind and $Y_n(t)$ is called Bessel function of the second kind of order n. The order n can be any real number, Bessel functions are expressible in series form. Bessel function of first kind is expressed as follows:

$$J_n(t) = t^n \sum_{r=0}^{\infty} \frac{(-1)^r (t/2)^{n+2r}}{2^n r! \Gamma(n + r + 1)} \tag{3}$$

For non-integer n, the functions $J_n(t)$ and $J_{-n}(t)$ are linearly independent, and are therefore the two solutions of the differential equation. On the other hand, for integer order n, the following relationship is valid i.e., $J_{-n}(t) = (-1)^n J_n(t)$ It can be readily shown that Bessel functions are orthogonal with respect to the weighting function t.

4.1 Fourier-Bessel Expansion for Speech Signal

The zeroth order Fourier-Bessel series expansion of a signal $x(t)$ considered over some arbitrary interval $(0, a)$ is expressed as [35]:

$$x(t) = \sum_{m=1}^{\infty} C_m J_0\Big(\frac{\lambda}{a}t\Big) \tag{4}$$

where $\lambda = \lambda_m, \quad m = 1, 2, 3, ...$ are the ascending order positive roots of $J_0(\lambda) = 0$ *and* $J_0\Big(\frac{\lambda_m}{a}t\Big)$ are the zero order Bessel functions. The roots of the Bessel function $J_0(\lambda)$ can be computed using the Newton-Raphson method. The sequence of Bessel functions $\{J_0\Big(\frac{\lambda_m}{a}t\Big)\}$ forms an orthogonal set on the interval $0 \leq t \leq a$ with respect to the weight t , that is

$$\int_0^a tJ_0\Big(\frac{\lambda_m}{a}t\Big)J_0\Big(\frac{\lambda_n}{a}t\Big)dt = 0, \quad for \quad m \neq n \tag{5}$$

using the orthogonality of the set $\{J_0\Big(\frac{\lambda_m}{a}t\Big)\}$, the FB coefficients C_m are computed by using the following equation

$$C_m = \frac{2\int_0^a tx(t)J_0\Big(\frac{\lambda_m}{a}t\Big)dt}{a^2[J_1(\lambda_m)]^2} \tag{6}$$

with $1 \leq m \leq Q$, where Q is the order of the FB expansion, and $J_1(\lambda_m)$ are the first order Bessel functions. The FB expansion order Q must be known apriori. The interval between successive zero-crossings of the Bessel function $J_0(\lambda)$ increases slowly with time and approaches π in the limit. If order Q is unknown, then in order to cover full signal bandwidth, the half of the sampling frequency and Q must be equal to the length of the signal.

The integral in the numerator of (6) is known as the finite Hankel transform (FHT). Many numerical computation methods are available to calculate the FHT and the corresponding FB coefficients [36–42]. It has been demonstrated in [43] that the order and range of non-zero coefficients of the FB series expansion of a test signal are changed as the center frequency and the bandwidth of the signal are varied. In particular that the range widens with larger bandwidth and the order increases with higher center frequency. There is one-to-one correspondence between the frequency content of the signal and the order (m) where the coefficient attains peak magnitude [44]. As the Fourier series coefficients are unique for a given signal, similarly FB series coefficients C_m are unique for a given

signal. However, unlike the sinusoidal basis functions in the Fourier series, the Bessel functions is having damped sinusoidal function and will decay over time. This features of the Bessel functions makes the FB series expansion suitable for non-stationary signals.

The initial value of the roots of Bessel function $J_0(m) = 0$ can be obtained by the following relation [45] $\lambda_{m+1} \approx \lambda_m + \pi, \quad 1 \leq m \leq Q$. In the next stage, each of the roots of the equation $J_0(m) = 0$ is determined accurately by the Newton-Raphson method in successive iteration. The iteration will be stopped when the root does not change its value significantly any more. The few roots are shown in Figure 2.

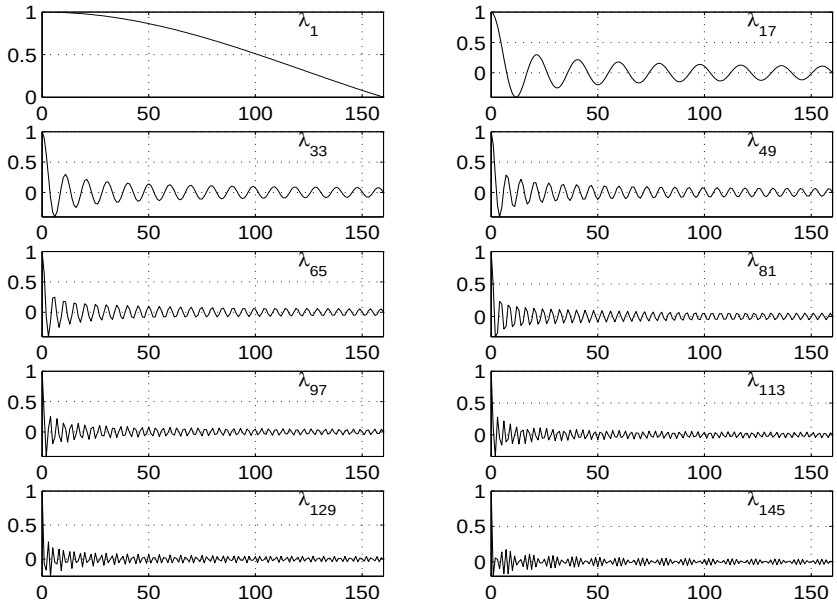

Fig. 2. *Bessel roots* $\{\lambda_m,\ m = 1, 2, ...,\ Q\}$ *are the ascending order positive roots of* $J_0(\lambda_m) = 0$.

The selection of optimum window size **a** is required for good resolution. A larger window provides a finer resolution in frequency, which also means that more number of FB coefficients will be needed to cover the same signal bandwidth. Since the FB coefficients are real, each signal component $x_i(t)$ can be directly reconstructed from the FB coefficient.

The synthesized signal from the FB coefficients has an inherent filtering property that helps to reduce the noise in low and high frequency regions of the spectrum. Noise reduction by the frequency domain techniques needs two different types of information processing, namely, magnitude and phase. The signal

can not be uniquely reconstructed without one of these quantities. The signal enhancement technique which use some kind of frequency domain processing, such as power spectrum subtraction, need the phase information to synthesize the signal [46] [47]. Since the FB decomposition is defined in the time-domain, there is no need for separate processing of the magnitude and phase information. Instead, the FB coefficients contain all the necessary information for the synthesis of the signal.

Moreover, the coefficients used to represent speech signal using Fourier-Bessel basis function are less when compared to sinusoidal basis function. For effective reconstruction of speech signal, very less number of FB coefficients is enough but this is not the case for Fourier series expansion. Bessel expansion has other properties like resolving multi-component signals without knowledge of frequency bands.

4.2 Analysis of Signal using FB Expansion

Considering multicomponent sinusoidal signal which has the frequencies at 500, 1000, and 2000 Hz, and its amplitude 10, 2, 0.5 respectively. FB coefficient of the multicomponent signal is obtained from the zeroth order FB function and the coefficients are plotted as shown in Figure 3. The original signal is reconstructed

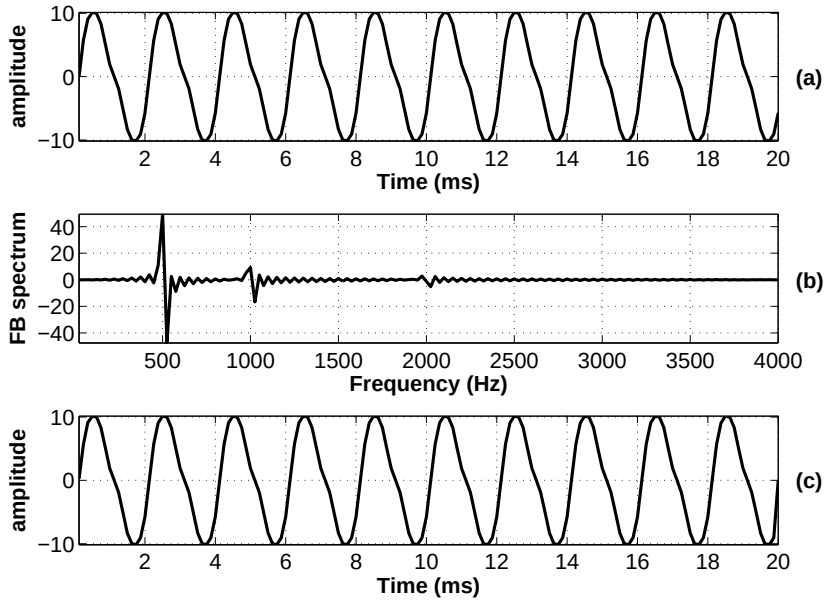

Fig. 3. *Multicomponent signal (a) original signal (b) FB coefficients (c) Reconstuction of the original signal using fewer number of FB coefficients.*

back by having only a fewer number of FB order that is considering first component FB coefficients (450:550), second component FB coefficients (950:1050) and third component FB coefficients (1950:2050)

Calculate the FB coefficient C_m for a given signal, every component of the multicomponent AM-FM signal has non-overlapping clusters of FB coefficients, each component is directly reconstructed from its FB coefficients.

4.3 Requisite Conditions for Fourier-Bessel Decomposition

Some requisite conditions for the Fourier-Bessel (FB) decomposition of a signal are listed below:

1. Since the FB decomposition can essentially represent only the oscillatory nature of a signal, the dc component of the signal, if any, should be removed prior to the decomposition.
2. The FB expansion order Q must be known a priori. Since, the interval between successive zero-crossing of the Bessel function $J_0(\lambda)$ increases with time and approaches π in the limit. If order Q is not known, then for covering full signal bandwidth, that is, the half of the sampling frequency, Q must be equal to the length of the signal.
3. The selection of the optimum window size **a** is required for effective separation of the components of a multicomponent signal. A larger window provides a finer resolution in frequency, which also means that more number of FB coefficients will be needed to cover the same signal bandwidth.

4.4 Applications and use of Fourier-Bessel Decomposition

1. A particular application where this method will be useful is in speech analysis, because speech can be modeled as a sum of AM and FM signals corresponding to formant frequencies, and one of the main objectives in the analysis of speech signals is to estimate the formant frequencies.
2. As Fourier-Bessel expansion uses the Hankel transform to calculate the FB coefficients, the shift variant property of the Hankel transform may prove valuable for non-stationary analysis.
3. The FB series based method for decomposition of a signal into its constituent components is advantageous over the technique based on the filter bank approach, because we do not need any prior information about the frequency-band of the signal.

4.5 Fourier-Bessel Cepstral Coefficients

The Fourier-Bessel cepstral coefficients (FBCC) are obtained by applying Mel filtering to Fourier-Bessel coefficients and taking the cepstrum or the Mel filtered coefficients, thereby characterizing the perceptual characteristics of human ear. The Mel filter bank is designed in the same way as in the case of MFCC by assigning frequencies to the index of the Fourier-Bessel coefficients.

The procedure for extraction of FBCC is similar to that of MFCC except for the fact that in FBCC we find the Fourier-Bessel coefficients instead of DFT coefficients. The block diagram for the computation of the FBCC is shown in Fig. 4 and procedure to extract FBCC is as follows.

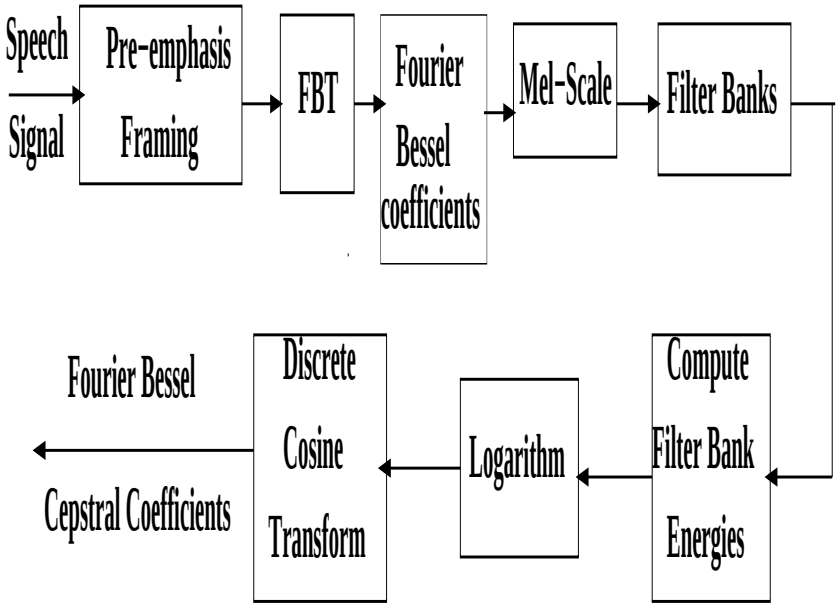

Fig. 4. *Block diagram illustrating the steps involved in the computation of the Fourier Bessel cepstral coefficients (FBCC).*

1 The preprocessing step like pre-emphasizing, DC offset removal and windowing are performed.
2 The Fourier-Bessel transform is applied for getting Fourier-Bessel coefficients for the windowed signal.
3 Mel filtering is done for the Fourier-Bessel coefficients to get Mel filtered Bessel coefficients.
4 Natural logarithm is taken on the absolute values of Mel filters Bessel coefficients to get log Mel filtered Bessel coefficients.
5 Discrete Cosine transform is applied on the log Mel filtered Bessel coefficients to get FBCC

The first and second derivatives of the time domain signal are also concatenated with the FBCC to get a larger dimensional feature vector. So 12 FBCC, 1 energy coefficient, 13 first and 13 second derivatives of FBCC to get a 39 dimensional feature vector for each frame of speech signal.

5 Experimental Evaluation

The evaluation of a speaker identification experiment was conducted in the following manner. The test speech was first processed by the front-end analysis to produce a sequence of feature vectors $\{\overrightarrow{x_1}, ..., \overrightarrow{x_t}\}$. To evaluate different test utterance lengths, the sequence of feature vectors was divided into overlapping segments of T feature vectors. The first two segments from a sequence would be

$$\overbrace{\{\overrightarrow{x_1}, \overrightarrow{x_2}, ..., \overrightarrow{x_T}\}}^{Segment1}, \quad \overrightarrow{x_{T+1}}, \overrightarrow{x_{T+2}}, ...$$

$$\overrightarrow{x_1}, \quad \overbrace{\{\overrightarrow{x_2}, ..., \overrightarrow{x_T}, \overrightarrow{x_{T+1}}\}}^{Segment2}, \quad \overrightarrow{x_{T+2}}, ...$$

For example a test segment length of 4 seconds corresponds to $T = 400$ feature vectors at a 10 ms frame rate. Each segment of T vectors was treated as a separate test utterance. The shift in frame was taken to be 1, (First system was tested with 1 to 400 feature vector, then with 2 to 401 and so on). The identified speaker of each segment was compared to the actual speaker of the test utterance and the number of segments which were correctly identified was tabulated.

$$\% \quad \text{correct identification} = \frac{\text{Number of correctly identified segments}}{\text{Total number of segments}} X100. \quad (7)$$

The process was repeated for all users and the average of these percentages was taken as the percentage recognition of the system.

Each speaker had approximately equal amounts of testing speech so the performance evaluation was not biased to any particular speaker. While there may be variations among the individual speakers performance, the aim of the evaluation measure was to track the average performance of the system for different speaker identification tasks, allowing common basis of comparison.

1. The experiment was performed on a closed set database of 20 speakers (10 male and 10 female). The database contained speech of 60 seconds for every speaker sampled at 16 kHz. To reduce the effect of noise and to improve the performance of system, preprocessing like pre-emphasis and DC offset removal was performed on the signal. The speech signal was segmented into frames of 20 ms length and frame shift is taken to be 10 ms. The MFCC and FBCC coefficients are computed for each frame separately. Twelve filter banks are taken in the frequency range of 0-8 kHz. After taking DCT, 11 coefficients (except the 0^{th} coefficients) are taken as MFCC coefficients and stored. Similarly 11 FBCC coefficients are calculated. The GMM system was trained with the first 30 seconds of the speech data available for each user. The covariance matrix type was set to be diagonal and the system was trained for GMM orders 2, 4 and 8. The next, 30 seconds of the speech of every speaker was used to test the system. The feature vectors are extracted

Number of Mixtures	SNR	10 seconds of training		30 seconds of training	
		% recognition for MFCC	% recognition for FBCC	% recognition for MFCC	% recognition for FBCC
2	10	9	20	17	18
	20	23	25	36	38
	30	54	55	72	74
	40	68	74	91	95
	50	78	80	96	96
	clean	80	82	97	98
4	10	12	13	17	30
	20	28	31	40	48
	30	57	59	77	79
	40	90	91	94	97
	50	91	92	99.65	99.73
	clean	96.21	96.61	99.49	99.86
8	10	20	22	29	37
	20	32	34	50	67
	30	58	62	85	93
	40	95	98	94	100
	50	97.51	98.95	100	100
	clean	97.68	99.26	100	100

Table 1. Performance evaluation of MFCC and FBCC over 10 second and 30 seconds of training from speech data of 60 seconds and testing was performed on a closed set database of 20 speakers (10 male and 10 female) using the remaining 30 seconds of speech data.

for the speech and the probability was calculated for each speaker. The speakers object for which the function returns the least value is taken as the correct speaker.

2. The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance. Corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT and verified and prepared for CD-ROM production by the National Institute of Standard Technology (NIST). From TIMIT database a random set of 50 speakers, 40 male and 10 female was taken for speaker identification system. In TIMIT database, 10 different utterance of 2-3 seconds of each are available for every speaker. Of these 10 utterances 5 are of SX type, 3 are of SI type and 2 are of SA type. The SX and SI type utterances concatenated to be used as training data, while the SA type utterances are concatenated to form testing data. In the process of feature extraction, for both MFCC

and FBCC, cepstral mean removal was done (i.e., after taking log of filter bank coefficients, normalization was performed).

Number of Mixtures	**SNR**	**20 sec of training and 5 sec of testing data**	
		% recognition for MFCC	% recognition for FBCC
2	10	42	43
	20	68	74
	30	90	91
	40	93	95
	50	93.89	95.45
	clean	94.23	96.02
4	10	40	52
	20	83	84
	30	91	93
	40	95	96
	50	95.21	97.20
	clean	95.48	97.24
8	10	45	49
	20	92	93
	30	95.59	98.50
	40	98.86	99.78
	50	98.78	99.42
	clean	100	100

Table 2. *Performance evaluation of MFCC and FBCC for a random set of 50 speakers (40 male and 10 female) from TIMIT database.*

Table 1 and Table 2 shows the performance of speaker identification for the 20 speakers (10 male 10 female) and random set of 50 speakers (40 male and 10 female) taken randomly from the TIMIT database.It can be seen from both the Tables 1 and 2 that FBCC gives better recognition accuracy than MFCC coefficients. Results from the Table 1 and 2 shows that recognition accuracy increases with increase in the order of GMM, this is due to the fact that the system can be modeled accurately by using higher order GMM. The FBCC outperforms MFCC significantly in noisy environment and the performance of the system depends on the training time for building the speaker model.

3. In TIMIT database of 630 speakers, of which 70% male and 30% female speakers speech utterances were taken for testing the speaker identification system. Among 10 different utterances of 2-3 seconds of each speaker 5 are of SX type, 3 are of SI type and 2 are of SA type. The SX and SI type utterances are concatenated to be used as training data, while the SA type utterances are used as testing data.
   The performance of LPCC, MFCC and FBCC features are considered for the application of speaker identification and is evaluated on TIMIT database.

The steps for the evaluation are as follows and the performance of these features for clean speech and different white noise SNR ratio is tabulated in Table 3.

- **Preprocessing:** To improve the performance of the system and to reduce the effects of noise. Preprocessing steps like pre-emphasis and DC offset removal was performed on the signal.
- **Framing and windowing:** The speech signal was split into frames of length 20 ms, the frame shift was taken to be 10 ms. Hamming window was applied to each frame to avoid abrupt discontinuities.
- **Feature extraction:** The LPCC, MFCC and FBCC coefficients are compared for each frame separately, 26 filter banks are taken in the frequency range of 0-8 kHz. After taking DCT, 12 coefficients (except the 0^{th} coefficient) are taken as MFCC coefficients. Similarly 12 FBCC coefficients are calculated.
- **Training the system:** The GMM system was trained with 5040 speech sentences that is among 630 speakers 8 utterances from each speaker is taken. The covariance matrix type was set to be *diagonal* and the system was trained for GMM orders 2, 4, 8, 16 and 32.
- **Testing:** The system was tested with 1260 speech utterances that is among 630 speakers 2 utterances per speaker is taken for testing the performance of the system. The feature vectors were extracted for the test speech and the probability was calculated for each speaker. The speaker object for which the function returns the maximum probability is taken as the correct speaker.
- **Performance evaluation:** The system was tested by taking 400 feature vectors at a time. The shift in frame was taken to be 1 sample (First, system was tested with 1 to 400 feature vectors, then next with 2 to 401 and so on). The percentage that the system identifies each speaker correctly was calculated. The process was repeated for all users and the average of these percentages was taken as the percentage recognition of the system.

6 Summary and Conclusion

The compact representation of speech is possible using Bessel functions because of similarity between voiced speech and the Bessel functions. Both voiced speech and Bessel functions exhibit quasi-periodicity and decaying amplitude with time. An important step in the speaker identification process is to extract sufficient information for good discrimination and at the same time, have to capture the information in a form and size that is amenable to effective modeling. The LPC spectral representation, such as LPC cepstral and reflection coefficients have been used extensively for speaker recognition; whereas these model based representation can be severely affected by noise. Recent studies have found that the directly computed filter bank features are more robust for noisy speech recognition. So the Fourier-Bessel cepstral coefficients (FBCC) are obtained by applying

SNR	Features	Number of Gaussian Mixtures				
		2	**4**	**8**	**16**	**32**
20 dB	LPCC	16.21	22.34	36.87	41.08	47.05
	MFCC	19.89	22.35	36.62	43.31	47.58
	FBCC	20.11	24.88	38.65	48.50	53.48
30 dB	LPCC	33.28	48.48	59.84	73.32	77.18
	MFCC	64.58	65.21	81.63	88.67	90.47
	FBCC	66.53	66.89	81.66	88.68	91.42
40 dB	LPCC	46.06	60.73	75.23	84.50	87.47
	MFCC	83.60	84.12	94.06	97.13	97.31
	FBCC	84.06	84.80	94.39	97.24	97.46
50 dB	LPCC	55.36	68.24	82.60	90.18	91.56
	MFCC	86.71	86.81	95.71	98.05	98.09
	FBCC	87.85	88.00	96.01	98.08	98.34
Clean	LPCC	59.72	71.47	84.07	92.53	93.10
	MFCC	87.52	88.52	96.27	98.17	98.38
	FBCC	88.49	88.91	96.52	98.22	98.38

Table 3. *Performance evaluation of LPCC, MFCC and FBCC over the TIMIT database for 630 speakers taking GMM orders of 2, 4, 8,16 and 32 for the white noise SNR ratios of 20, 30, 40, 50 db and clean speech data.*

Mel filtering to Fourier-Bessel (FB) coefficients and taking the cepstrum or the Mel filtered coefficients. We have build the speaker models from the FB features derived from the speech samples, as an alternative to Mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC). Evaluated the performance of LPCC, MFCC and FBCC features using the Gaussian mixture models on the TIMIT database which consists of 630 speakers and 10 speech utterances per speaker. From the resulted tables it is observed that FBCC outperforms MFCC and LPCC. Recognition accuracy increases with increase in the order of GMM. This is due to the fact that the system can be modeled accurately by using higher order GMM. The performance of the system depends on the training time for building the speaker model. Fourier-Bessel cepstral features (FBCC) are likely to be more robust to noise, hence the effects of environment on the systems performance can be reduced.

References

1. J. Gonzaler Rodriguez, J. Ortega Garcia, and J. J. Lucena Molina: On the application of the Bayesian approach to real forensic conditions with GMM-based systems. A speaker Odyssey- the speaker recognition workshop (2001) 135–138
2. M. Sigmund: Speaker recognition identifying people by their voices. Habilitation thesis, Brno University of Technology, Institute of radio electronics (2000)
3. G. R. Doddington: Speaker recognition identifying people from their voices. Proc. IEEE **73** (1985) 1651–1664
4. Douglas O'Shaughnessy: Speaker recognition. IEEE Acoust. Speech and Signal Process. Mag. **3(4)** (1986) 4–17
5. D. A. Reynolds, Thomas F. Quatueri, and Robert B. Dunn,: Speaker Verification using Adapted Gaussian Mixture Models,. Digital Signal Processing **10, no. 1** (2000) 19–41

6. J. R. Deller, Jr. J. G. Proakis, and J. H. L. Hansen: Discrete-Time Processing of Speech Signals. Macmillan, New York (1993)
7. B. S. Atal: Automatic recognition of speakers from their voices. Proc. IEEE **64** (1976) 460–475
8. A. E. Rosenberg: Automatic speaker verification: A review. Proc. IEEE **64 (4)** (1976) 475–487
9. S. Furui: An overview of speaker recognition technology. In Automatic Speech and Speaker recognition (1996) 31–56
10. D. O. Shaughnessy: Speech communication: Human and Machine. Digital Signal Processing, Reading MA: Addison-Wesley, New York (1987)
11. Rabiner L. and Juang B. H.: Fundamentals of speech recognition. Pearson Education. First Indian Reprint (2003)
12. R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue: Survey of the state of the art in Human Language Technology. Cambridge University Press (1997)
13. J. Tierney: A study of LPC analysis of speech in additive noise. IEEE Trans. Acoust. Speech Signal Processing **ASSP-28** (1980) 389–397
14. S. S. Stevens and J. Volkman: The relation of pitch to frequency. American J. Psych. **53 (3)** (1940) 329–353
15. S. B. Davis and P. Mermelstein: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Processing **28** (1980) 357–366
16. R. Bracewell: The Fourier Transform and its Applications. McGraw-Hill (1964)
17. L. Dolansky: Choice of base signals in speech signal analysis. IRE Trans. on Audio **AU-8** (1960) 221–229
18. H. J. Manley: Analysis-synthesis of connected speech in terms of orthogonalized exponentially damped sinusoids. J. Acoust. Soc. Am. **35** (1963) 464–474
19. C. Chen, K. Gopalan, and P. Mitra: Speech signal analysis and synthesis via Fourier-Bessel representation. IEEE Trans. Acoust. Speech Signal Processing **10** (1985) 497–500
20. F. S. Gurgen and C. S Chen: Speech enhancement by Fourier-Bessel coefficients of speech and noise. IEE Proc. **137, no. 5** (1990) 290–294
21. C. S. Chen and K. S. Huo: Loeve method for data compression and speech synthesis. IEE Proc.-I **138, no. 5** (1991)
22. K. Gopalan, T. R. Anderson, and E. J. Cupples: A comparision of speaker identification result using features based on cepstrum and Fourier-Bessel expansion. IEEE Trans. Speech and Audio Processing **7 (3)** (1999) 289–294
23. K. Gopalan: Speech coding using Fourier-Bessel expansion of speech signals. In Proc. 27^{th} Annu. Conf. IEE Industrial Electronics Society **3** (2001) 2199–2203
24. J. Schroeder: Signal processing via Fourier-Bessel series expansion. Digital Signal Processing **3** (1993) 112–124
25. A. Papoulis: Signal Analysis. McGraw-Hill N. Y. (1977)
26. I. H. Sheddon: Fourier Transforms. McGraw-Hill N. Y. (1951)
27. Rose R. C. and Reynolds D. A.: Text-independent speaker identification using automatic acoustic segmentation. In Proc. Inter. Conf. Acous. speech and Signal Processing (ICASSP) (1990) 293–296
28. D. A. Reynolds: A Gaussian mixture modeling approach to text-independent speaker identification. PhD. Thesis, Georgia Institute of Technology (1992)
29. D. A. Reynolds and R. C. Rose: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Processing **3** (1995) 72–83
30. D. A. Reynolds: Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. **17** (1995) 91–108
31. D. A. Reynolds: Automatic speaker recognition using Gaussian mixture speaker models. Lincoln Lab. J. 8 (1996) 173–192

32. Doddungton G., Przybocki M., Martin A., and D. A. Reynolds: The NIST speaker recognition evaluation-overview, methodology, systems, results, perspectives. Speech Commun. **31 (2-3)** (2000) 225–254
33. Martin A. and Przybocki M.: The NIST 1999 speaker recognition evaluation-an overview. Digital Signal Processing (2000) 1–18
34. D. A. Reynolds: Comparision of background normalization methods for text-independent speaker verification. In Proc. European Conf. Speech Processing and Technology (1997) 963–966
35. D. Slepian, H. J. Landau, and H. O. Pollack: Prolate spheroidal wave functions, Fourier analysis and uncertainity principle I and II . Bell system Technical Journal **40, no.1** (1961) 43–84
36. A. V. Oppenheim, G. V. Frisk, and D. R. Martinez: An algorithm for the numerical evaluation of the Hankel transform. Proc. IEEE **66** (1978) 264–265
37. E. Cavanagh and B. Cook: Numerical evaluation of Hankel transforms via Gaussian Lagurerre polynomial expansions. IEEE Trans. Acoust. Speech Signal Processing **ASSP28** (1979) 361–366
38. S. M. Candel: An algorithm for the Fourier-Bessel transform . Comp. Physics Communication **23** (1981) 343–353
39. S. M. Candel: Dual algorithms for fast calculation of the Fourier-Bessel transform. IEEE Trans. Acoust. Speech Signal Processing **ASSP-29** (1981) 963–972
40. S. M. Candel: Fast computation of Fourier-Bessel transform . In Proc. Inter. Conf. Acoust. Speech and Signal Processing (ICASSP) **3** (1982) 2076–2079
41. K. Gopalan and C. S. Chen : Numerical evaluation of Fourier-Bessel expansion . In Proc. Inter. Conf. Acoust. Speech and Signal Processing (ICASSP) (1983) 14–16
42. A. V. Oppenheim, G. V. Frisk, and D. R. Martinez: Computation of the Hankel transform using projections . J. Acoust. Soc. Am. **68** (1980) 523–529
43. A. Potamianos and P. Maragos: A comparision of the energy operator and Hilbert transform approaches for signal and speech demodulation. Signal Processing **37, no. 1** (1994) 95–120
44. R. B. Pachori and P. Sircar: EEG signal analysis using FB expansion and second-order linear TVAR process. Signal Processing **88, no. 2** (2008) 415–420
45. : United States National Burean of Standards Computation Laboratory. Tables of Bessel functions of fractional order Columbia university press NEWYORK (1948)
46. J. S. Lim: Speech Enhancement. (Prentice Hall N. J.)
47. S. F. Boll: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Processing **26** (1978) 471–472